Sliced Inverse Regression for datastreams: An introduction

Stéphane Girard

Sliced Inverse Regression for datastreams: An introduction

Stéphane Girard

Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Abstract

In this tutorial, we focus on data arriving sequentially by block in a stream. A semiparametric regression model involving a common EDR (Effective Dimension Reduction) direction β is assumed in each block. Our goal is to estimate this direction at each arrival of a new block. A simple direct approach consists of pooling all the observed blocks and estimating the EDR direction by the SIR (Sliced Inverse Regression) method. But in practice, some disadvantages become apparent such as the storage of the blocks and the running time for high dimensional data. To overcome these drawbacks, we propose an adaptive SIR estimator of β. The proposed approach is faster both in terms of computational complexity and running time, and provides data storage benefits. A graphical tool is provided in order to detect changes in the underlying model such as a drift in the EDR direction or aberrant blocks in the data stream. This is a joint work with Marie Chavent, Vanessa Kuentz-Simonet, Benoit Liquet, Thi Mong Ngoc Nguyen and Jérôme Saracco.

1 Sliced Inverse Regression (SIR)

1.1 Multivariate regression

Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$. The goal is to estimate $G : \mathbb{R}^p \to \mathbb{R}$ such that

$$Y = G(X) + \xi \quad \text{where } \xi \text{ is independent of } X.$$

- Unrealistic when p is large (*curse of dimensionality*).
- **Dimension reduction** : Replace X by its projection on a subspace of lower dimension without loss of information on the distribution of Y given X.
- **Central subspace** : smallest subspace S such that, conditionally on the projection of X on S, Y and X are independent.

1.2 Dimension reduction

- Assume (for the sake of simplicity) that $\dim(S) = 1$ *i.e.* S =span(b), with $b \in \mathbb{R}^p \Longrightarrow$ **Single index model:**

$$Y = g(< b, X >) + \xi$$

  where ξ is independent of X.

- The estimation of the p-variate function G is replaced by the estimation of the univariate function g and of the direction b.

- **Goal of SIR** [Li, 1991] : Estimate a basis of the central subspace. (*i.e.* b *in this particular case.*)

1.3 SIR

Idea:

- Find the direction b such that $< b, X >$ best explains Y.

- Conversely, when Y is fixed, $< b, X >$ should not vary.

- Find the direction b minimizing the variations of $< b, X >$ given Y.

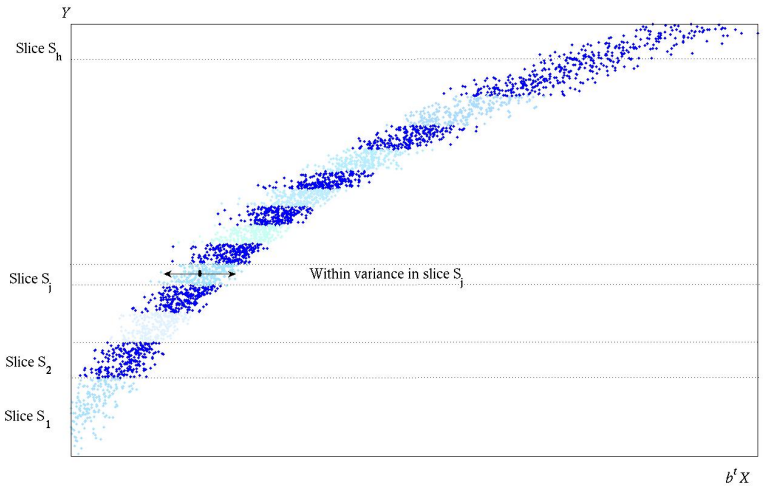
In practice:

- The support of Y is divided into h slices S_j.

- Minimization of the within-slice variance of $< b, X >$ under the constraint $\text{var}(< b, X >) = 1$.

- Equivalent to maximizing the between-slice variance under the same constraint.

1.4 Illustration



1.5 Estimation procedure

Given a sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, the direction b is estimated by

$$\hat{b} = \underset{b}{\operatorname{argmax}}\, b'\hat{\Gamma}b \;\text{ such that }\; b'\hat{\Sigma}b = 1. \tag{1}$$

where $\hat{\Sigma}$ is the empirical covariance matrix and $\hat{\Gamma}$ is the between-slice covariance matrix defined by

$$\hat{\Gamma} = \sum_{j=1}^{h} \frac{n_j}{n}(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})', \quad \bar{X}_j = \frac{1}{n_j}\sum_{Y_i \in S_j} X_i,$$

where n_j is the number of observations in the slice S_j.
The optimization problem (1) has a closed-form solution: $\hat{b}$ is the eigenvector of $\hat{\Sigma}^{-1}\hat{\Gamma}$ associated to the largest eigenvalue.

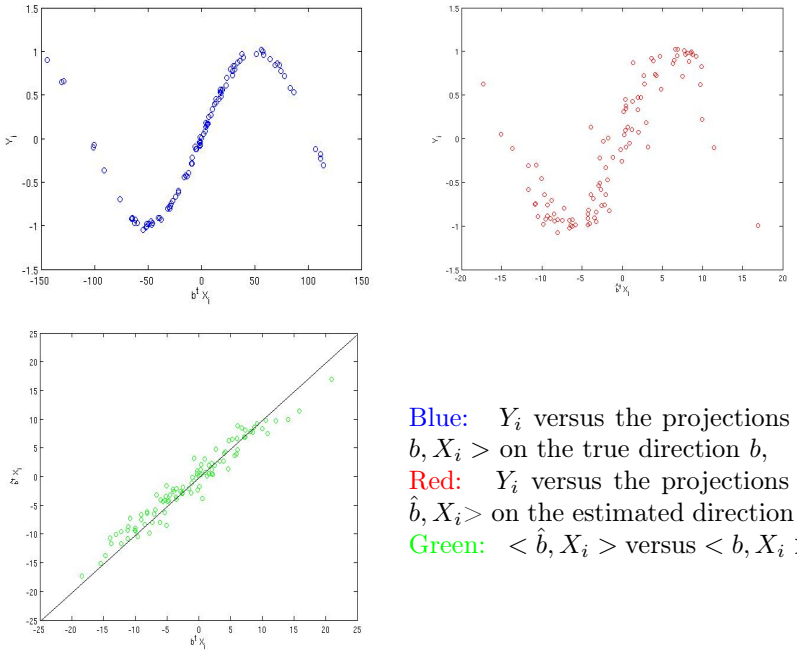
1.6 Illustration

Simulated data.

- Sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size $n = 100$ with $X_i \in \mathbb{R}^p$, dimension $p = 10$ and $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$.
- $X_i \sim \mathcal{N}_p(0, \Sigma)$ where $\Sigma = Q\Delta Q'$ with
  - Δ =diag$(p^2, \ldots, 2^2, 1^2)$,

- Q is an orientation matrix drawn from the uniform distribution on the set of orthogonal matrices.

- $Y_i = g(< b, X_i >) + \xi_i$ where
  - g is the link function $g(t) = \sin(\pi t/2)$,
  - b is the true direction $b = 5^{-1/2}Q(1, 1, 1, 1, 1, 0, \ldots, 0)'$,
  - $\xi \sim \mathcal{N}_1(0, 9.10^{-4})$



Blue: Y_i versus the projections $< b, X_i >$ on the true direction b,
Red: Y_i versus the projections $< \hat{b}, X_i >$ on the estimated direction $\hat{b}$,
Green: $< \hat{b}, X_i >$ versus $< b, X_i >$.

2 SIR for data streams

2.1 Context

- We consider data arriving sequentially by blocks in a stream.
- Each data block $t = 1, \ldots, T$ is an i.i.d. sample (X_i, Y_i), $i = 1, \ldots, n$ from the regression model $Y = g(< b, X >) + \xi$.
- **Goal**: Update the estimation of the direction b at each arrival of a new block of observations.

2.2 Method

- Compute the **individual directions** $\hat{b}_t$ on each block $t = 1, \ldots, T$ using SIR.

- Compute a **common direction** as

$$\hat{b} = \underset{||b||=1}{\operatorname{argmax}} \sum_{t=1}^{T} \cos^2(\hat{b}_t, b) \cos^2(\hat{b}_t, \hat{b}_T).$$

Idea: If $\hat{b}_t$ is close to $\hat{b}_T$ then $\hat{b}$ should be close to $\hat{b}_t$.
Explicit solution: $\hat{b}$ is the eigenvector associated to the largest eigenvalue of

$$M_T = \sum_{t=1}^{T} \hat{b}_t \hat{b}_t' \cos^2(\hat{b}_t, \hat{b}_T).$$

2.3 Advantages of SIRdatastream

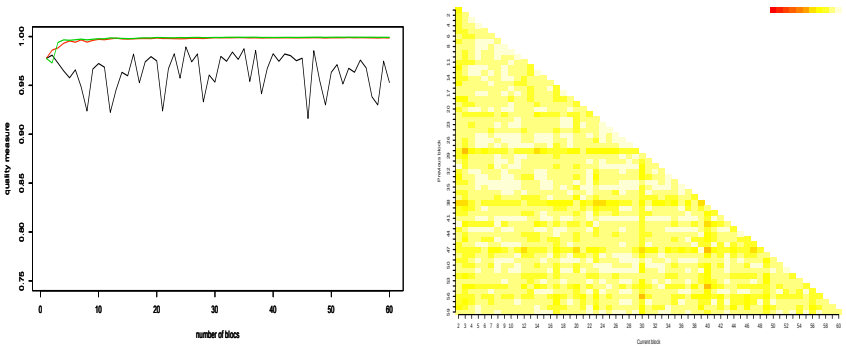
- Computational complexity $O(Tnp^2)$ v.s. $O(T^2np^2)$ for the brute-force method which would consist in applying regularized SIR on the union of the t first blocks for $t = 1, \ldots, T$.
- Data storage $O(np)$ v.s. $O(Tnp)$ for the brute-force method.

(under the assumption $n >> \max(T, p)$).

- Interpretation of the weights $\cos^2(\hat{b}_t, \hat{b}_T)$.

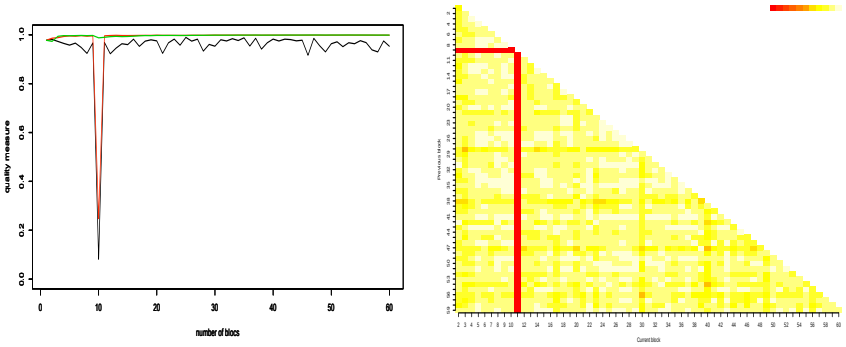
2.4 Illustration on simulations

2.4.1 Scenario 1: A common direction in all the 60 blocks.



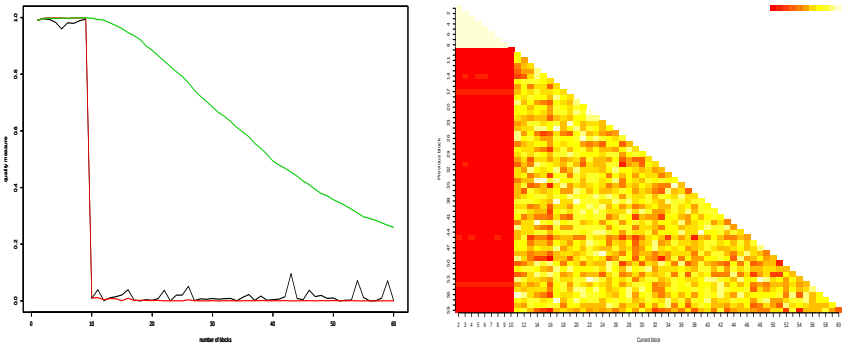
Left: $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right:* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.

2.4.2 Scenario 2: The 10th block is different from the other ones.

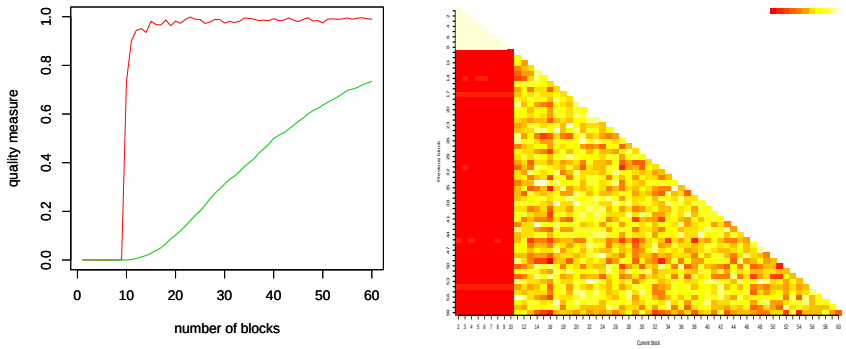


Left: $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right:* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.

2.4.3 Scenario 3: A drift occurs from the 10th block (b to $\tilde{b}$)

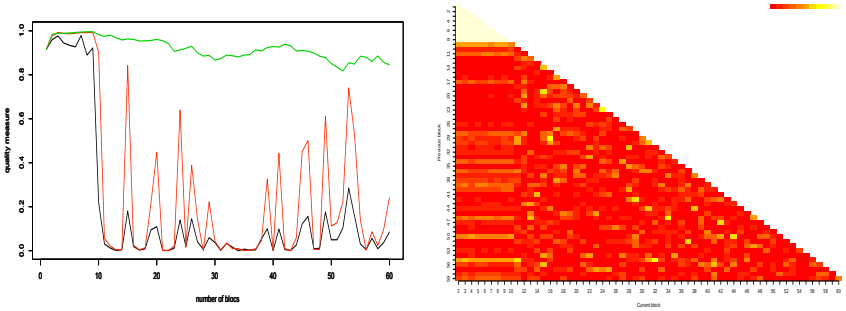


Left: $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right:* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.



Left: $\cos^2(\hat{b}, \tilde{b})$ for SIRdatastream and SIR brute-force. *Right:* $\cos^2(\hat{b}, \tilde{b})$

2.4.4 Scenario 4: From the 10th block to the last one, there is no common direction.



Left: $\cos^2(\hat{b}, b)$ for SIRdatastream, SIR brute-force and SIR estimators at each time t. *Right:* $\cos^2(\hat{b}_t, \hat{b}_T)$. The lighter (yellow) is the color, the larger is the weight. Red color stands for very small squared cosines.

3 Application to real data

3.1 Estimation of Mars surface physical properties from hyperspectral images

Context:

- Observation of the south pole of Mars at the end of summer, collected during orbit 61 by the French imaging spectrometer OMEGA on board Mars Express Mission.
- 3D image: On each pixel, a spectra containing $p = 184$ wavelengths is recorded.

- This portion of Mars mainly contains water ice, CO_2 and dust.

Goal: For each spectra $X \in \mathbb{R}^p$, estimate the corresponding physical parameter $Y \in \mathbb{R}$ (grain size of CO_2).

3.2 An inverse problem

Forward problem.

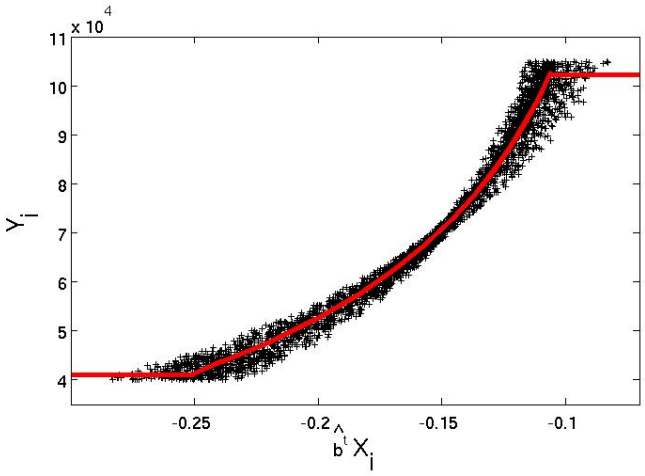
- Physical modeling of individual spectra with a surface reflectance model.
- Starting from a physical parameter Y, simulate $X = F(Y)$.
- Generation of $n = 12,000$ synthetic spectra with the corresponding parameters.

$\Longrightarrow$ Learning database.

Inverse problem.

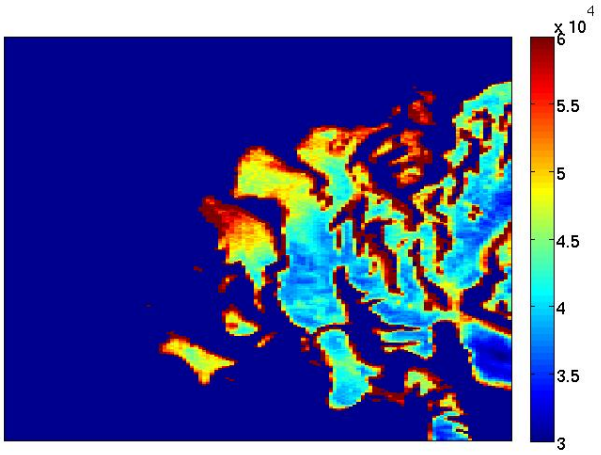
- Estimate the functional relationship $Y = G(X)$.
- Dimension reduction assumption $G(X) = g(< b, X >)$.
- b is estimated by SIR, g is estimated by a nonparametric one-dimensional regression.

3.3 Estimated function g



Estimated function g between the projected spectra $< \hat{b}, X >$ on the first axis of SIR and Y, the grain size of CO_2.

3.4 Estimated CO_2 maps



Grain size of CO_2 estimated with SIR on a hyperspectral image of Mars.

References

[1] Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86, 316–327.

[2] Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. & Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets*, 114, E06005.

[3] M. Chavent, S. Girard, V. Kuentz, B. Liquet, T.M.N. Nguyen & J. Saracco. A sliced inverse regression approach for data stream, *Computational Statistics*, 29, 1129–1152.

[4] A. Chiancone, F. Forbes & S. Girard. Student Sliced Inverse Regression, *Computational Statistics and Data Analysis*, 113, 441–456, 2017.

[5] A. Chiancone, S. Girard & J. Chanussot. Collaborative Sliced Inverse Regression, *Communications in Statistics - Theory and Methods*, 46, 6035–6053, 2017.

[6] S. Girard & J. Saracco. *An introduction to dimension reduction in nonparametric kernel regression*, In D. Fraix-Burnet and D. Valls-Gabaud, editors, Regression methods for astrophysics, volume 66, pages 167–196, EDP Sciences, 2014.

[7] R. Coudret, S. Girard, & J. Saracco. A new sliced inverse regression method for multivariate response, *Computational Statistics and Data Analysis*, 77, 285–299, 2014.

[8] C. Bernard-Michel, L. Gardes & S. Girard. A Note on Sliced Inverse Regression with Regularizations, *Biometrics*, 64, 982–986, 2008.

[9] C. Bernard-Michel, L. Gardes & S. Girard. Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, 19, 85–98, 2009.

[10] A. Gannoun, S. Girard, C. Guinot & J. Saracco. Sliced Inverse Regression in reference curves estimation, *Computational Statistics and Data Analysis*, 46(3):103-122, 2004.

[11] Barreda, L., Gannoun, A., Saracco, J. (2007). Some extensions of multivariate SIR. *Journal of Statistical Computation and Simulation*, 77(1-2), 1-17.

[12] Barrios, M.P.; Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statist. Probab. Lett.*, 77(3), 247-255.

[13] Chavent, M., Kuentz, V., Liquet, B. and Saracco, J. (2011). A sliced inverse regression approach for a stratified population. *Communications in statistics - Theory and methods*, 40, 1-22.

[14] Chen, C-H and Li, K-C (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8, no. 2, 289-316.

[15] Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, 22, 1-26.

[16] Duan, N., Li, K.C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, 19, 505-530.

[17] Liquet, B., Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics - Simulation and Computation*, 37(6), 1198-1218.

[18] Liquet, B. and Saracco, J. (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Comput. Stat.*, 27, 103-125.